Shuffle and Learn: Unsupervised Learning using Temporal Order Verification

Ishan Misra C. Lawrence Zitnick Martial Hebert

Carnegie Mellon University, The Robotics Institute

Facebook AI Research

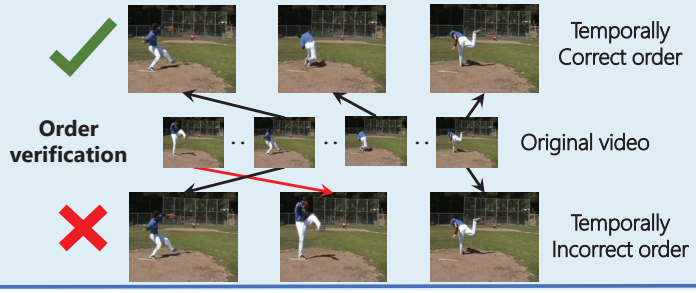
Overview

Goal

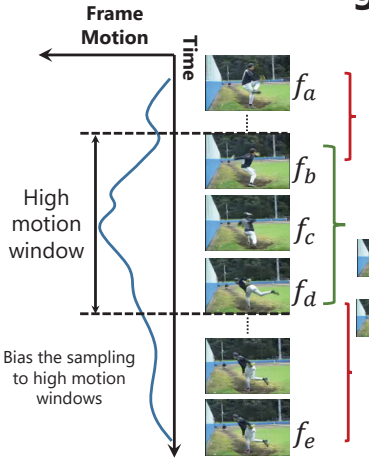
- Use raw spatiotemporal signal from videos to learn a visual representation
- No supervised information for learning.
- Little pre-processing of input video
- Application to action recognition and pose estimation

Key Ideas

- Extract sequence of frames from video
- Either leave frames in their original temporal order, or shuffle them.
- Learning task: Verify if frames are in correct temporal order

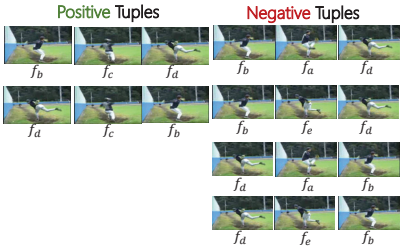


Mining Tuples

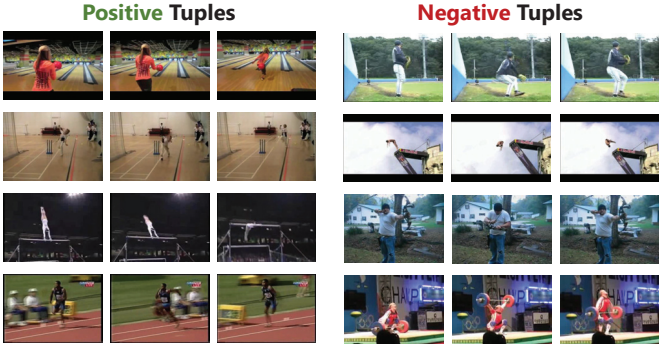


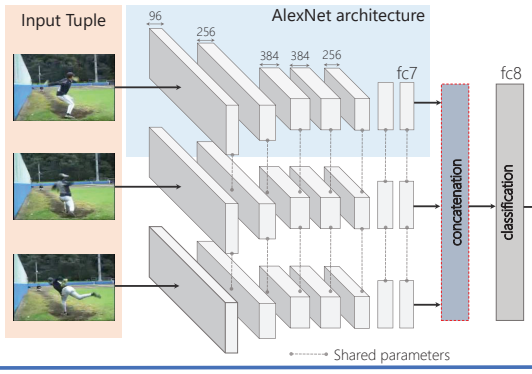
1. Compute Frame Level Optical Flow
2. Find high motion temporal window
3. Bias sampling using optical flow
4. Sample positive and negative tuples such that:

$$\min(|f_a - f_b|, |f_e - f_d|) > \tau_{min}$$
$$|f_b - f_d| < \tau_{max}$$



Approach





- Mine ~900k tuples from UCF101
- Initialize randomly
- Use batch-normalization
- Train for 10 epochs
- Binary Cross-entropy loss

Correct order Or Not

$$\text{Loss} = \sum_i p_i y_i + (1 - p_i)(1 - y_i)$$

Minimize using SGD over all the tuples

Action Recognition Results

Ablation Analysis

τ_{max}	τ_{min}	Tuple	Action
30	15	60.2	47.2
60	15	**72.1**	**50.9**
60	60	64.3	49.1

% neg	% pos	Tuple	Action
50	50	52.1	38.1
75	25	**72.1**	**50.9**
85	15	67.7	48.6

Dataset	Initialization	Mean Classification Accuracy
UCF101	Random	38.6
	Ours	50.2
	ImageNet pre-trained	**67.1**
HMDB51	Random	13.3
	Ours	18.1
	UCF101 pre-trained	15.2
	ImageNet pre-trained	28.5
	ImageNet + Ours	**29.9**

Comparison against existing unsupervised methods

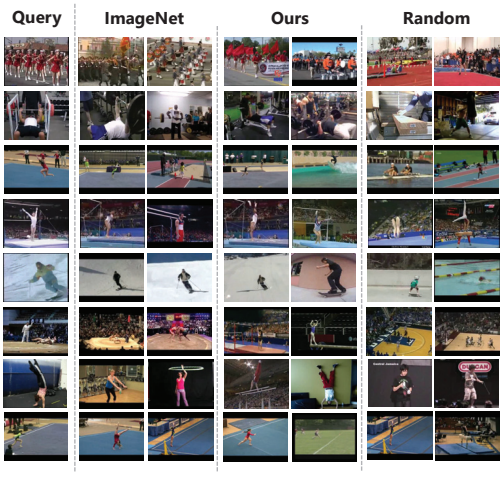
Dataset	Two Close	Two Order	DrLim	TempCoh	Video Patch	Ours
UCF101	42.3	44.1	45.7	45.4	40.7	**50.9**
HMDB51	15.0	16.4	16.3	15.9	15.6	**19.8**

DrLim: Hadsell et al., 2006; TempCoh: Mobahi et al., 2009; VideoPatch: Wang and Gupta, 2015

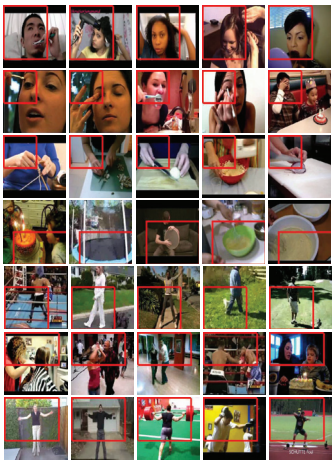
UCF101: 101 Action classes; ~9k train, ~1k test
HMDB51: 51 Action classes; ~3.4k train, 1.4k test

Qualitative Results

Nearest Neighbors from Different Videos



Visualizing pool5 units



Pose Estimation Results

Measuring PCK and PCKh on keypoint estimation.
Upper body for FLIC; Full body for MPII

Initialization	FLIC Dataset Mean PCK	FLIC Dataset AUC PCK	MPII Dataset Mean PCKh@0.5	MPII Dataset AUC PCKh@0.5
Random	74.5	36.1	72.9	34.0
Video Patch	77.1	42.1	82.8	43.8
DrLim	65.2	27.9	84.3	41.5
Ours	**84.9**	**49.6**	**87.7**	**47.6**
UCF pre-train	78.8	42.0	86.9	45.5
ImageNet pre-train	85.8	51.3	85.1	47.2
ImageNet+Ours	**86.2**	**52.5**	**88.0**	**49.5**

Conclusion

- Needs little pre-processing and no human supervision
- Learns a powerful visual representation for human pose
  - Correct way to answer if frames are shuffled is to focus on humans
  - At par with supervised methods
- Complementary information to supervised representations

Code and Models: https://git.io/vPkht